The Use of Tests in the Evaluation of Methods of Instruction

BUFORD JOHNSON

Bureau of Educational Experiments, New York City

Results of Investigations at the Johns Hopkins University Summer School.

THE fundamental purpose underlying the origination and development of educational measurements was their use for the improvement of instruction. We seem scarcely to have passed from the first stage of developing a measure that may be considered standard, and certainly have gone but a little way beyond the evalution of the status of system, class, or individual, in comparison with the obtained standard.

We believe the most effective use of measurements is in the determination of methods of instruction, meaning neither specific devices and forms of procedure, nor grouping according to intelligence ratings, but an elasticity of program that will make possible the adjustment to individual needs of drill, assignments, forms of recitation and groupings of activities.

The pupils in the five grades, 4th to 8th inclusive, of the Demonstration School of the Johns Hopkins University Summer Course for 1918 were children who had failed of promotion because of deficiency in one or more fundamental subjects. There were two exceptions—children who had not failed but wished to attend the summer session. These deficiencies were usually marked arithmetic or English, the latter including reading, spelling, composition and grammar.

From this composition we would expect the distribution curves for standardized tests in these subjects to be skewed toward the lower end.

Standard tests were given to these pupils by the students in the university course of experimental education, under direct supervision of the instructor. Records as to age and deficiencies were taken from the cards sent out by individual instructors in the various schools of the city from which the pupils were drawn. Each of the teachers of these five classes was asked during the

last week to rate her pupils in accordance with a judgment scale worked out by the Psychological Survey of New York. We will consider the records of these scores and ratings as a basis for determining methodology of instruction.

The following tests were used: The Cleveland Survey Arithmetic Tests; Woody Arithmetic Scales—Subtraction and Division, Series A; Stone Reasoning Tests; Starch Grammar Tests including only Parts of Speech, Nouns and Pronouns, Punctuation Scale and Grammatical Scale; Kansas Silent Reading Tests; Boston Copying Tests: Psychological Survey Judgment Scale.

The question of over-ageness contributes little to the causative factors. Table I shows the distribution of deficiencies according to the age and grade of the pupils. Numbers in the columns designated by curriculum subjects represent the frequency of deficiencies in those subjects. In the 7th grade there was one pupil fifteen years old and one sixteen years old, both marked "deficient" in arithmetic. In the 8th grade there was one pupil sixteen and one seventeen years old. The latter was a cripple and had an Intelligence Quotient of 80.2, determined by the Stanford Revision of the Binet-Simon Scale. The mode for each grade falls at the expected age.

Approximately, 60 per cent. of the pupils in the 4th, 6th and 8th grades; 75 per cent. of the 5th grade; and $16^2/_3$ per cent. of the 7th grade were marked as deficient in both arithmetic and English. The table comparing the central tendency of the group with tentative norms established indicates normal conditions in some subjects, especially for the 6th grade who are above in all tests taken, but slightly below in judgment rating, while in other subjects the 7th and 8th grades are below the standard in varying degrees. These are shown in Table 1A. The greatest deficiency shown is in the grammar tests. This indicates a general low standing of the class, while variations in other subjects could easily be explained by a few outstanding cases.

Tables 2 to 7 inclusive show the distribution of scores in each test. There is a marked clustering about the norms except in the grammar tests. Superficially, this would indicate the expected standing for all except a very few cases in arithmetic, reading, punctuation and simple grammatical usage. But we believe that those in the central group and above, even of the exceptionally

TABLE 1
AGE—DEFICIENCY DISTRIBUTION

Grade	Age	Arithmetic and English	English	Arithmetic	No Deficiency	Total
IV	9	1		1		2
	10	5		7		12
	11	3		4		7
	12	1		1		2
	13	3	3			6
	not given	1	1	?		4
	Total	14	4	15		33
V	10			1		1
	11	8	1	3		12
	12	5	1			6
	13	5		1		6
	14	3				3
	not given	1				1
	Total	22	2	5		29
VI	No ages given	18	2	17		37
VII	12			4		4
	13	2	2	7		12
	14	4	2	9		15
	15			5		5
	16			1		1
	Total	6	5	26		37
VIII	12		1			1
	13	1	2		1	4
	14	8	4	1		13
	15	9	3	3		15
	16	1	1		1	3
	17	1				1
	Total	20	11	4	2	37

TABLE 1 A

Comparison of Central Tendencies with Established Norms.*

	Grade VI %	*Grade VII* %	*Grade VIII* %
Woody-Division,	+13	+ 9	—10
Cleveland Survey Arithmetic,		— 8	— 8
Stone Reasoning,	+ 6	+13	— 3
Kansas Silent Reading,	+20	— 7	— 2
Parts of Speech,		—39	—40
Nouns and Pronouns,		—28	—19
Grammatical,		0	— 4
Punctuation,		0	— 4
Boston Copying,	+ 9	—21	+ 6
Judgment Rating,	— 4	+13	+34

*The figures of the table represent the per cents of the established norms by which the classes exceed or fall short of the norms.

high standing, should have an analysis of the errors made. Some erroneous habit, some lack of speed or accuracy must have been at the basis of the teacher's estimate in selecting such cases for the extra Summer Session.

TABLE 2

WOODY ARITHMETIC SCALES, SERIES A—SUBTRACTION FOR GRADES IV, V AND VI—DIVISION FOR GRADES VII AND VIII—NUMBER IN EACH GRADE THAT SOLVED THE LAST 20 PROBLEMS CORRECTLY

Grade	SUBTRACTION			DIVISION	
	IV	V	VI	VII	VIII
Example No. 16, ...	29	24	32	32	24
" " 17, ...	30	21	34	31	24
" " 18, ...	25	22	32	22	14
" " 19, ...	25	19	31	30	21
" " 20, ...	22	18	32	28	22
" " 21, ...	12	19	31	31	21
" " 22, ...	18	18	32	31	23
" " 23, ...	18	16	29	31	19
" " 24, ...	12	19	31	23	20
" " 25, ...	9	15	26	25	14
" " 26, ...	0	1	27	26	18
" " 27, ...	0	1	24	29	24
" " 28, ...	0	7	25	31	22
" " 29, ...	1	10	13	29	18
" " 30, ...	0	7	23	23	14
" " 31, ...	1	9	19	21	15
" " 32, ...	0	0	14	24	15
" " 33, ...	2	9	16	19	19
" " 34, ...	2	9	12	23	17
" " 35, ...	3	11	12	8	6
" " 36, ...	0	0	0	2	0
Median, ...	21.33	26.5	27.8	29	23.3
Class Score, ...	5.69	7.25	7.38	7.17	6.43
Standard Score, ...	4.22	5.47	6.46	6.59	7.16
Total Number Pupils, ...	34	24	35	34	27

TABLE 3

CLEVELAND SURVEY SERIES OF ARITHMETIC TESTS—COMPARISON OF MEDIAN NUMBER OF PROBLEMS SOLVED BY SEVENTH AND EIGHTH GRADES OF JOHNS HOPKINS UNIVERSITY DEMONSTRATION SCHOOL WITH THOSE FOR CLEVELAND SCHOOLS

CLEVELAND TEST	EIGHTH GRADE	CLEVELAND EIGHTH	SEVENTH GRADE	CLEVELAND SEVENTH
A	32	27.5	28	26.7
B	25	26	22	21.5
C	19	19	17	17.7
D	22	22.5	22	20.8
E	7	7.8	6	7.5
F	8	10.1	10	8.6
G	5	6.6	5	5.9
H	6	8.5	0	7.7
I	4	4.7	4	4.0
J	4	5.7	4	4.9
K	11	12.5	11	10.1
L	3	3.9	3	3.2
M	4	5.1	3	4.4
N	2	2.6	3	2
O	4	5.5	3	4.1

TABLE 4

STONE REASONING TESTS—DISTRIBUTION OF SCORES FOR SIXTH, SEVENTH AND EIGHTH GRADES.

	VI	VII	VIII
Score 3.1— 4,	3		1
" 4.1— 5,	2		0
" 5.1— 6,	4	5	1
" 6.1— 7,	8	4	3
" 7.1— 8,	4	8	5
" 8.1— 9,	2	4	3
" 9.1—10,	1	3	3
" 10.1—11,	3	1	0
" 11.1—12,	1	2	2
" 12.1—13,	2	1	0
" 13.1—14,	1	2	3
" 14.1—15,	0	3	0
" 15.1—16,	1		1
" 16.1—17,			1
Total,	32	33	23
Median,	6.96	8	8 5
Score attained by Upper 80% of Class,	5.2	6.4	6.6
Standard Score,	6.5	7.5	8.75
Percentage that reached Standard,	70	70	52
Percentage of Accuracy,	77.7	78	80.7
Standard of Accuracy,	80	85	90

Under the routine adopted, even in the summer session, the necessary stimulation for the correction of the trouble will not be given. The teachers considered many quite up to standard, as the judgments show, and one remarked that she thought some of the boys just did not take interest in the regular school work, though they certainly had the ability to do it. Only by an elaboration of testing, the scheme of which is built upon the analysis of the individual performance, can the instruction be devised that makes for conservation of energy and interest.

From the standpoint of the class needs, the percentage solving correctly certain problems may be considered. Fifty-three per cent. of the seventh grade solved correctly all except the last three examples in the Woody Test, only two pupils solving the last example. The time limit enters into the evaluation of the degree of difficulty of these problems at the end. Fifty-six per cent. of the eighth grade solved 32 of the 36 problems, with no one getting the last one. The following examples are found to be troublesome for both the seventh and eighth grades:

$2 \div 2$

$\frac{5}{4} \div \frac{3}{5}$

$13\overline{)65065}$

$9\frac{5}{8} \div 3\frac{3}{4}$

$75\overline{)2250300}$

$248 \div 7$

$2400\overline{)504000}$

$25\overline{)9750}$

$12\overline{)2.76}$

$23\overline{)469}$

$\frac{3}{4} \div 5$

$3\frac{1}{2} \div 9$

Four of these and the last three examples of the test are included in the list of troublesome ones found in the Janesville, Wisconsin, Educational Survey. These examples may easily be analyzed into specific abilities necessary for automatic use of the processes and combinations. Does common usage make it more difficult to grasp the division of a number by itself? We ordinarily say, "Give the two children one each," or "There is one apiece for the five." The second and third problems listed belong to a second type, involving zeros that are often puzzling, but when the operation is mastered it is usually of permanence and effectiveness. The fourth problem lends itself to a short cut that should be easily acquired. Decimals and common fractions present difficulties widely variant according to individual needs. Certainly an analysis of these specific errors should be made and met by constructive planning.

In the Copying Tests spelling of words, omitted words and undotted "i's" were decidedly the most frequent errors; many other forms, such as capitalization and punctuation, require no more than the ordinary practice given to such. Grammar Tests show the greatest difficulty with parts of speech, nouns and pronouns.

TABLE 5
KANSAS SILENT READING TEST—DISTRIBUTION OF SCORES

SCORE	GRADE				
	IV	V	VI	VII	VIII
Scores falling between 0 and .9,					
" " " 1 " 1.9,					
" " " 2 " 2.9,	1				
" " " 3 " 3.9,	2				
" " " 4 " 4.9,					
" " " 5 " 5.9,					
" " " 6 " 6.9,	4	4	2		
" " " 7 " 8.9,	3	7	3		1
" " " 9 " 10.9,	5	4	1	6	3
" " " 11 " 12.9,	7	1	4	3	1
" " " 13 " 14.9,	4	3	5	5	3
" " " 15 " 17.9,	7	2	7	8	4
" " " 18 " 20.9,	1		2	5	12
" " " 21 " 23.9,		1	5	3	1
" " " 24 " 26.9,			4		1
" " " 27 " 29.9,			1	2	
" " " 30 " 34.9,		1	1	2	1
" " " 35 " 39.9,		1			
" " " 40 " 44.9,					
Total number of Pupils,	34	24	35	34	27
Median Score,	11.1	9.6	16.1	15.4	18.4
Standard Score,	9.9	13.7	13.4	16.5	18.8
Twenty-five Percentile,	7.0	7.3	11.5	11.8	14.5
Median Score,	11.1	9.6	16.1	15.4	18.4
Seventy-five Percentile,	13.8	13.4	22.3	20.1	20.3

TABLE 6
STARCH GRAMMAR TESTS—DISTRIBUTION OF SCORES FOR SEVENTH AND EIGHTH GRADES

Test	No. 1		No. 2			Punctuation Scale		Grammatical Scale	
	Grade		Grade			Grade		Grade	
Score	VII	VIII	VII	VIII	Score	VII	VIII	VII	VIII
0.0 — 4.9	0	0	5	2	0.0 — 0.9	1	1		
5.0 — 9.9	8	1	10	4	1.0 — 1.9	1			
10.0 — 14.9	8	4	4	8	2.0 — 2.9				
15.0 — 19.9	2	2	5	3	3.0 — 3.9				
20.0 — 24.9	5	6	3		4.0 — 4.9				
25.0 — 29.9	4	3	1	1	5.0 — 5.9				
30.0 — 34.9	3	1	1		6.0 — 6.9	1	2	1	1
35.0 — 39.9					7.0 — 7.9	5	3	7	4
40.0 — 44.9	1				8.0 — 8.9	14	4	5	5
					9.0 — 9.9	4	5	3	1
					10.0 — 10.9	2	2	6	2
					11.0 — 11.9		1	1	
Total Number of pupils,	26	17	29	18		28	18	23	18
Median,	18.5	20	9	11.5		8	8	8	8
Standard Score,	30	33	13	16		8	8.3	8	8.3
Average,	19.3	19.5	11.7	11.2		7.5	7.8	8.4	8

TABLE 7
BOSTON COPYING TEST—DISTRIBUTION OF ERRORS

Errors	GRADE IV	V	VI	VII	VIII	Total No. Errors
Spelling,	70	30	46	64	20	230
Capitalization,	39	23	16	38	17	133
Omitted words,	128	78	49	72	37	364
Added Words,	9	3	3	9	3	27
Wrong Words,	17	28	25	22	22	114
Punctuation,	8	4	5	3	4	24
Undotted "i's",	41	3	72	99	80	295
Uncrossed "t's",	15	1	21	18	3	58
Misplaced Words,	*5	0	*14	0	0	19
Total Number of Errors,	332	170	251	320	186	1259

*By one pupil.

TABLE 8
TEACHER'S JUDGMENT SCALE—DEVISED BY PSYCHOLOGY SURVEY OF NEW YORK

GRADE	IV					V					VI					VII					VIII				
Score	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Section I,	7	15	9	4	0	0	9	11	0	0	0	19	13	6	3	0	10	9	6	12	0	1	6	14	11
" II,	10	8	8	7	2	0	10	10	0	0	4	25	4	6	2	2	11	14	6	4	0	0	6	16	10
" III,	11	9	8	4	3	0	7	9	4	0	0	11	23	4	3	1	4	11	18	3	0	4	4	13	11
" IV,	5	12	10	5	3	0	7	12	1	0	0	11	18	9	3	1	5	14	9	8	0	3	13	11	5
" V,	5	17	6	4	3	0	9	9	2	0	0	15	10	1	5	0	12	8	4	13	0	3	13	11	5
Total Frequencies,	28	61	41	24	11	0	42	51	7	0	4	81	78	26	16	4	42	56	43	40	0	11	42	65	42
Average Rating,	12.4					13.3					14.2					16.9					19.3				

Section I—
ABILITY TO MAKE GOOD PERSONAL IMPRESSION. The combined impression made by Facial Expression, Bearing, Speech, Voice, Personal Neatness, Cheerfulness, Courtesy.

Section II—
ABILITY TO RESPOND TO TASKS PRESENTED BY THE TEACHER. Give attention to what is being taught; Become interested in his lessons; Acquire facts; Use his knowledge.

Section III—
ABILITY TO ADAPT CONDITIONS OF SCHOOL WORK. Be punctual; Follow the program; Take responsibility; Be orderly; Do continuous work.

Section IV—
ABILITY TO FIT INTO THE SCHOOL GROUP. Hold his own; Be companionable; Stand up for school interests; Gain respect.

Section V—
PROMISE OF ACHIEVING FUTURE SUCCESS. Ability to think quickly and arrive at the correct conclusion; Initiative; Steadfastness of purpose; Ability to get along with and influence others; Determination to succeed.

All children who rank highest in the group to be judged in the characteristics listed for a section are scored 5. Lowest score 1 should include those showing marked deficiency. If the class represents an unselected group of pupils, more than three-quarters are expected to fall into the three ranks, High (4), Middle (3) and Low (2).

When we follow up the individual records, we find some who are marked deficient only in arithmetic making a satisfactory score in the standard tests in arithmetic, but showing greater deficiency in grammar tests. The seven making highest scores in grammar tests, ranking above the standard score, are all marked deficient in English. This may mean a failure in spelling or composition, in which they are not tested; but it also means that the teacher for the summer instruction did not know the real difficulties so that the situation could be adjusted for the individual's specific needs.

The scores of the individual pupils show more strikingly the wide range of abilities or difficulties. (See Table 9.) The standard norms for the respective grades in the subjects listed are taken as the basis of the computation. The percentage above or below this standard of the individual score is given in the table. The individuals are chosen to represent the extremes rather than the average. Individual A had the highest average ranking of the 7th grade; N the lowest. In each case, however, the contrasting specific disability or ability is very marked. Individual L represents the highest group of the 8th grade. His rating in the tests in which he failed to make the standard score places him not far below the norm. The scores of the 7th grade boy in the various series of the Cleveland Survey Tests (See Table 10) illustrates well the need of analysis of performances. Since this arithmetic test is of the so-called spiral formation, the various fundamental operations enter into it by intervals and with increasing complexity. The span of attention, the speed of performance and the instability in certain automatic processes can be measured more successfully,

TABLE IX

Comparison of Individual Scores with Standard.

	Individual A Best in Gr. VII %	*Individual N Poorest in Gr. VII* %	*Individual L Best in Gr. VIII* %
Woody-Division, ...	0	0	+14
Stone Reasoning, ...	0	—30	+55
Kansas Silent Reading, ...	+25		
Parts of Speech, ...	+40	—57	+18
Nouns and Pronouns, ...	+24	—54	—10
Grammatical, ...	—25	+24	—16
Punctuation, ...	0	—14	— 3
Boston Copying, ...	—25	—12	+ 7
Judgment Rating, ...	+14	—17	

thereby. In multiplication this boy never attained a high score, while in subtraction he is consistently capable. Failure in division in no sense describes his status so far as that process is considered. Why he should have failed in the simpler forms, but succeeded in the more complex forms is a problem for study. Addition offers a similar situation. Such presentations emphasize the necessity of diagnosing specific disabilities.

TABLE X

Performance of a Boy in Grade VII in the Cleveland Survey Arithmetic Tests.

		% of Norm.
A	Addition,	+36
B	Subtraction,	+36
C	Multiplication,	+ 5
D	Division,	— 7
E	Addition,	—34
F	Subtraction,	+19
G	Multiplication,	— 8
H	Fractions,	— 5
I	Division,	—14
J	Addition,	+ 5
K	Division,	+29
L	Multiplication,	—21
M	Addition,	—28
N	Division,	+ 9

An analysis of the records in comparison with standard scores suggests three groupings. One group includes those who have accomplished what is usually expected of children in such grades. This is shown by the following percentages of the pupils in the 6th, 7th and 8th grades who have attained the norms in the specified subjects.

	ARITHMETIC	GRAMMAR	READING
Grade VI, ...	59%	.....	70%
Grade VII, ...	60%	41%	58%
Grade VIII, ...	44%	36%	70%

They need more varied applications to cultivate rapidity and efficient use of the fundamental operations as tools.

The second group, just below the standard score or class median, have specific difficulties, many of which a constructive program, based on the diagnosis made from the tests, could remedy, with economy of effort and time on part of pupils and teacher.

A third group have individual problems. In this class would fall 5 in arithmetic in the 6th grade; 6 in arithmetic, 11 in gram-

mar in the 7th grade: 7 in arithmetic and 7 in grammar in the 8th grade. Some of these doubtless have a much slower rate of learning. A detailed study of their needs would suggest the remedial measures.

The opportunity for instruction during a special session, independent of regular group routine, is an advance step for a system in meeting such individual failures. This should be an opportunity for remedying specific defects rather than marking time. The most significant result of the use of the tests comes from a study in detail of the facts secured, a more accurate rating of pupils, and a reconstruction in method on the basis of the facts obtained.